DisMix: Generating Order-Aware Mixup Augmentations for Robust Medical Image Classification

Dileepa Pitawela
AIML, Adelaide University
Adelaide, Australia

Gustavo Carneiro
CVSSP, University of Surrey
Surrey, UK

Hsiang-Ting Chen
AIML, Adelaide University
Adelaide, Australia

{dileepa.pitawela,tim.chen}@adelaide.edu.au g.carneiro@surrey.ac.uk

Abstract

Image mixup is a widely adopted data augmentation strategy, yet it is ill-suited for ordinal classification tasks such as medical disease grading, where labels encode a progression of severity. By indiscriminately blending disease-severity cues (ordinal) with appearance-level variation (non-ordinal), standard mixup produces samples that distort the very ordinal structure that underpins clinical severity grading. We introduce DisMix, an order-aware mixup framework for ordinal classification. DisMix disentangles ordinal and non-ordinal features via a dual-codebook VQ-VAE, allowing each subspace to be mixed independently: ordinal codes are interpolated to produce meaningful intermediate ranks, while non-ordinal codes are varied to introduce appearance diversity without corrupting the ordinal signal. Across four medical imaging datasets, DisMix shows the best aggregate performance among six image mixup baselines paired with six ordinal classifiers and remains robust under data scarcity and clinical grading variability.

1 Introduction

Image mixup augmentation is a widely used technique for training robust deep neural networks, improving generalization by generating intermediate samples that encourage smoother decision boundaries [13]. However, image mixup strategies [57, 27, 23, 41] are primarily developed for natural-image classification and are poorly suited for ordinal classification tasks such as medical imaging. In this domain, labels commonly follow an ordinal structure, reflecting a graded progression of disease severity rather than discrete, mutually independent categories. When conventional mixup is applied directly in this context, these methods indiscriminately blend ordinal cues (e.g., lesion severity) with non-ordinal visual factors such as imaging artifacts, background, or staining differences. This entangled mixing corrupts the very structure that ordinal labels are intended to capture. By ignoring the underlying features that carry the ordinal signal, the mixing can dilute or override rank-defining cues, producing samples with biologically implausible structures that deviate from the intended ordinal progression.

Fig. 1 illustrates this failure mode and motivation for DisMix. In the original Knee Osteoarthritis (KOA) dataset [8], samples with the same rank organize into distinct clusters. After naive mixing, this structure deteriorates: samples from different ranks become entangled, intermediate points drift away from the ordinal trajectory, and the overall feature distribution loses its ordered structure. Rather than enriching the data distribution, these mixed samples distort it, producing anatomically inconsistent samples (e.g., disoriented knee joint gap) that may hinder the learning of robust ordinal classifiers. In contrast, mixing using disentangled latents produces biologically plausible and order consistent samples with DisMix.

GenAI for Health Workshop @ NeurIPS 2026, Sydney.

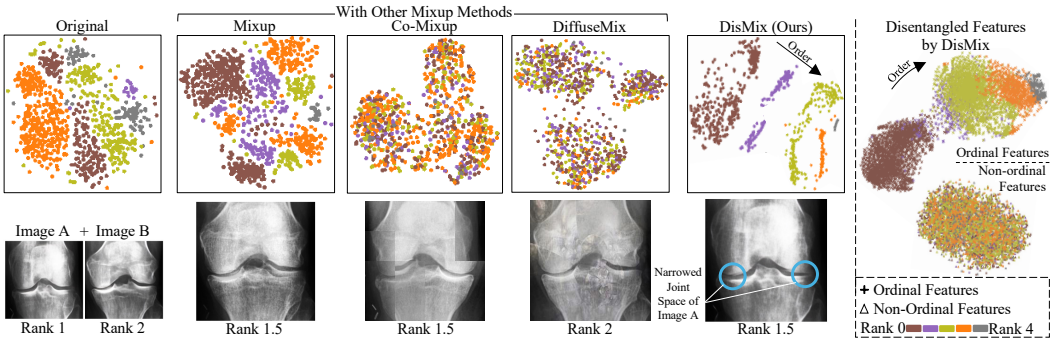


Figure 1: **Main panel (left):** Original feature space of the Knee Osteoarthritis (KOA) dataset [8], features and images from Mixup [57], Co-Mixup [31], DiffuseMix [23], and our DisMix. **Top row:** Existing mixup entangles features and collapses the structure of data that ordinal labels are meant to capture. DisMix preserves clusters of original data and enforces ordinal structure. **Bottom row:** Naive mixing produces distorted images or anatomically inconsistent structures (e.g., disoriented knee joint gaps), whereas mixing using disentangled latents, DisMix produces plausible, order consistent samples. **Rightmost panel:** DisMix disentangles ordinal and non-ordinal factors of data into separate latent spaces, learning an ordered manifold for former and a rank-invariant cluster for latter. (t-SNEs were produced by passing mixup images through an ILSVRC pretrained ResNet-50 [51]).

We address the fundamental incompatibility between conventional mixup and ordinal learning, recognizing that not all feature mixing is consistent with the underlying rank progression and that order-aware mixing is therefore necessary. To this end, DisMix trains a dual-codebook VQ-VAE that factorizes each image into an ordinal latent capturing rank-defining structure and a non-ordinal latent encoding rank-independent variations. The separation is enforced using soft ordinal supervision on the ordinal branch and an adversarial rank-removal objective on the non-ordinal branch. Building on this factorization, we propose a set of order-aware mixup policies (Tab. 1) that operate separately on the two subspaces, generating order-consistent augmented samples without corrupting the underlying ordinal signal. To summarize,

- We introduce DisMix, an order-aware mixup method that disentangles ordinal and non-ordinal features into distinct latent subspaces, enabling synthesis of order-consistent augmented samples.
- We propose a set of order-aware mixup policies that operate in the disentangled latent space, allowing controlled mixing during downstream ordinal classifier training.
- We comprehensively evaluate DisMix on four medical imaging datasets, against six mixup baselines using six ordinal classifiers.

Across 24 mixup-classifier combinations, DisMix achieves the best accuracy and MAE in 20 and 15 settings, respectively. Compared to the strongest mixup baseline, this consistent advantage yields statistically significant aggregate accuracy gains (one-sided Wilcoxon signed-rank test, $p = 0.0075$) and MAE reductions of 4–6%, while remaining robust under data scarcity and grading variability.

2 Related Work

Ordinal Classification. Ordinal classification preserve class order using specialized objectives, as standard multi-class losses ignore the relationships between adjacent ranks. ORCNN decomposes prediction into cumulative binary thresholds [40]; CNNPOR and MWR enforce pairwise and local-to-global ordinal constraints [38, 47]; POE models boundary uncertainty through probabilistic embeddings [37]; Ord2Seq recasts ranking as a sequence prediction [53]; and GOL learns geometry-aware ordinal representations [36]. To improve robustness to grading variability, SORD adopts soft labels [11], RNC aligns embeddings with rank order via contrastive learning [56], and CLOC learns adaptive margins between neighboring ranks [42]. While robust to mild grading variability,

Code: https://github.com/dpitawela/DisMix

these methods do not explicitly address the grader disagreements common in medical imaging, which typically occur between adjacent ranks [42], especially when ordinal cues are entangled with appearance variations. Conventional image mixup worsens this by indiscriminately blending ordinal and non-ordinal features. We instead disentangle these factors and apply controlled mixup: label-preserving by mixing non-ordinal factors and label-interpolating by mixing ordinal ones, improving supervision under grading variability.

Disentangled Representation Learning. Disentangled representation learning aims to separate data into latent factors that capture distinct sources of variation. VQ-VAEs introduce discrete codebooks promoting organized latent representations [50], with conditional VQ-VAE, FactorQVAE, and QLAE extending this paradigm through conditional guidance, total-correlation regularization, and improved semantic organization [63, 2, 20, 21]. Adversarial methods disentangle factors using mutual information, attribute supervision, or pretrained generators [9, 46, 29, 54]. Architectural designs further promote factorized representations [28, 35, 30], enabling applications such as anatomy/pathology separation, pose- and age-invariant face recognition, content-style decomposition, and ordinal content preservation [33, 49, 19, 62, 22, 59]. Diffusion models explore feature factorization [26, 34], but aim high-quality generation at substantially higher computational cost. In contrast, we use relatively light-weight VQ-VAE and disentangle ordinal and non-ordinal features from class labels alone, without attribute supervision, and perform order-aware latent mixup.

Image Mixing Augmentation. Image mixing augments data by combining samples and labels; however, existing methods are largely semantics-agnostic on which factors are being mixed. Vanilla Mixup linearly blends pixels and labels [57]; CutMix pastes image patches and mixes labels by area [55]; PuzzleMix, Co-Mixup, and GuidedMixup preserve salient regions during mixing [32, 31, 27]; and Manifold Mixup performs interpolation in latent space [52]. RankMixup uses rank-informed labels [41]; SUMix learns uncertainty-aware soft labels [44]; MOM smooths ordinal targets for manifold mixup [1]; and SGD-Mix and DiffuseMix propose label-preserving augmentation [12, 23], but none explicitly separate ordinal and non-ordinal features for mixing. In ordinal settings, labels reflect ordinal features, while non-ordinal variations are expected to be irrelevant. Prior methods mix entangled latents, either corrupting labels or under-utilizing augmentation by ignoring appearance cues. Our method instead disentangles ordinal and non-ordinal features, enabling controlled mixing in separate latent subspaces for improved label fidelity and semantic control.

Grading Variability. Medical datasets frequently exhibit grading variability across assessments due to inherent ambiguity in clinical interpretation [5, 25]. Existing approaches address this through majority voting, probabilistic consensus models [10, 48, 15], or by jointly estimating annotator reliability during training [4, 17, 18]. However, these methods primarily estimate consensus, have seen limited study in ordinal settings [60, 6], and have not been explored for order-preserving latent mixup. The closest related work, CLOC [42], learns ordinal representations under grading variability but assumes a single assessment. In contrast, our method focuses on preserving ordinal geometry through order-aware latent mixup under grading variability.

3 Preliminaries

Ordinal classification. Ordinal classification is a special case of multi-class classification where classes follow a natural ordering denoted by $r_0 \prec r_1 \prec ... \prec r_{C-1}$, where C is the number of classes, and $\prec$ indicates the ranking relation. Ordinal classification assumes that images are composed of ordinal features, which determine the rank, and rank-invariant non-ordinal features [59]. Beyond accuracy, ordinal consistency is measured by mean absolute error (MAE) accounting for the magnitude of rank errors. Throughout this paper, the terms class, label, grade, and rank are used interchangeably, while grading variability refers to disagreements among annotators.

Problem Setup. Let the training set be $\mathcal{D}=\{(x_i, \{y_{i,a}\}_{a\in\mathcal{A}})\}_{i=1}^{N}$, where each sample $x_i \in \mathcal{X}$ has a set of ordinal labels $y_{i,a} \in \mathcal{Y}=\{0, 1, ..., C-1\}$ that reflect grading variability among annotators $\mathcal{A}$.

Ordinal Soft Labels. Standard soft labels capture annotation uncertainty but ignore ordinal relationships between classes. Ordinal soft labels capture both uncertainty and ordinal relation. For

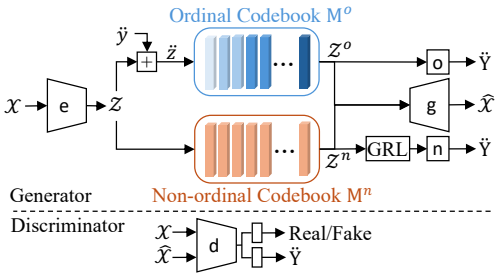


Figure 2: Overview of DisMix architecture, featuring a dual-codebook VQ-VAE generator for modeling ordinal and non-ordinal feature spaces, and a multi-task discriminator for enforcing ordinal consistency and perceptual quality.

Table 1: Latent mixing and labels assignment in mixup policies. The mixed sample is obtained by decoding $\tilde{z}^o$ and $\tilde{z}^n$.

Policy	Latent Interpolation	Label Assignment
Ordinal Mix	$\tilde{z}^o = \lambda z_i^o + (1-\lambda) z_j^o$, $\tilde{z}^n \in \{z_i^n, z_j^n\}$	$\tilde{y} = \lambda \bar{y}_i + (1-\lambda)\bar{y}_j$
Non-Ordinal Mix	$\tilde{z}^o \in \{z_i^o, z_j^o\}$, $\tilde{z}^n = \lambda z_i^n + (1-\lambda) z_j^n$	$\tilde{y} = \bar{y}$ of sample providing $\tilde{z}^o$
Order Swap	$(\tilde{z}^o, \tilde{z}^n) \in \{(z_j^o, z_i^n), (z_i^o, z_j^n)\}$	$\tilde{y} = \bar{y}$ of sample providing $\tilde{z}^o$
Generate & Mix	$\tilde{z}^o = \lambda z^{o,\text{anc}} + (1-\lambda) z^{o,x'}$, $\tilde{z}^n \in \{z^{n,\text{anc}}, z^{n,x'}\}$	$\tilde{y} = \lambda \bar{y}^{\text{anc}} + (1-\lambda) y'$

each x_i with scalar ranks $\{y_{i,a}\}_{a\in\mathcal{A}}$, each rank is first transformed into a progressive binary vector. These vectors are then averaged across annotators to compute the ordinal soft target $\ddot{y}_i$ with $\ddot{y}_i^{(k)} = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \mathbb{I}(k<y_{i,a})$, for $k \in \{0, ..., C-2\}$, where $\ddot{y}_i \in \ddot{\mathcal{Y}} \subset \mathbb{R}^{C-1}$ and $\mathbb{I}(\cdot)$ is an indicator function. Examples are provided in App. A.

Ordinal supervision is provided via the Soft Ordinal Regression (SOR) loss [11], computed as binary cross-entropy over ordinal soft labels, which penalizes prediction errors in proportion to their ordinal distance from the target during DisMix generator training.

4 Methodology

Conventional mixing of ordinal images disrupts the underlying ordinal structure by indiscriminately blending rank-defining ordinal and rank-independent non-ordinal features; DisMix instead learns separate latent subspaces for each, enabling independent interpolation of ordinal features to generate intermediate ranks and non-ordinal features to introduce appearance diversity without disrupting ordinal semantics.

4.1 Disentangling Ordinal and Non-Ordinal Factors

DisMix employs a dual-codebook VQ-VAE (Fig. 2) to disentangle ordinal and non-ordinal features. Formally, the encoder $\mathsf{e} : \mathcal{X} \to \mathcal{Z}$ maps an input image x to a latent feature map z, which is then factorized into an ordinal component $z^{\mathrm{o}} \in \mathcal{Z}^{\mathrm{o}}$ and a non-ordinal component $z^{\mathrm{n}} \in \mathcal{Z}^{\mathrm{n}}$ via the ordinal and non-ordinal branches. Each branch maintains a dedicated discrete codebook to facilitate a compact representation. The decoder $\mathsf{g} : \mathcal{Z}^{\mathrm{o}} \times \mathcal{Z}^{\mathrm{n}} \to \mathcal{X}$ reconstructs the image from the quantized codes $\hat{x} = \mathsf{g}(z^{\mathrm{o}}, z^{\mathrm{n}})$, enabling the model to learn meaningful decompositions while preserving visual fidelity. The framework is optimized jointly with an adversarial discriminator $\mathsf{d} : \mathcal{X} \to \{0, 1\} \times \ddot{\mathcal{Y}}$, which ensures perceptual realism and ordinal consistency.

Ordinal Quantizer. The ordinal branch maintains a learnable codebook $\mathbf{M}^o \in \mathbb{R}^{|\mathcal{Z}|\times D^o}$, with D^o vectors of dimension $|\mathcal{Z}|$, to capture ordinal features. To bias towards ordinal semantics, the encoder features are conditioned on the ordinal soft label $\ddot{y}$. Specifically, the label is projected and added to the encoder output, followed by sigmoid gating to emphasize rank-relevant channels: $\ddot{z} = (z + h_1(\ddot{y})) \odot h_2(h_1(\ddot{y}))$, where $h_1 : \ddot{\mathcal{Y}} \to \mathcal{Z}$, $h_2 : \mathcal{Z} \to [0, 1]^{|\mathcal{Z}|}$, and $\odot$ denotes elementwise multiplication.

Instead of the standard distance-based lookup, we adopt attention-based quantization [7], allowing the conditioned features to attend selectively to codebook entries by treating $\ddot{z}$ as *queries* and the codebook entries as *keys* and *values*. To reduce over-smoothing of ordinal features, we replace soft attention with hard attention: $z^o = \mathsf{one_hot}\left(\mathrm{softmax}(\ddot{z}^\top \mathbf{M}^o)\right)(\mathbf{M}^o)^\top$, where $\mathsf{one_hot}(\cdot)$ selects the highest-attention codebook entry. During training, hard assignments are used in the forward pass, with gradients propagated through the soft attention weights using the straight-through estimator.

To encourage the codebook to learn order-discriminative features, we attach a linear classifier $\mathsf{o} : \mathcal{Z}^o \to \ddot{\mathcal{Y}}$ that predicts rank from z^o. This auxiliary supervision is imposed via the SOR loss $\ell_{\mathrm{C}}^{\mathrm{o}}(z^o, \ddot{y}) = \mathrm{SOR}(\mathsf{o}(z^o), \ddot{y})$ to align quantized embeddings with the ordinal structure of data.

Non-Ordinal Quantizer. The non-ordinal branch maintains a learnable codebook $\mathbf{M}^n \in \mathbb{R}^{|\mathcal{Z}| \times D^n}$ paired with an adversarial classifier to capture order-invariant features. Given the encoder embedding z, the nearest codebook entry is selected to obtain the quantized representation z^n.

To suppress ordinal information from this branch, we attach a linear classifier $\mathsf{n} : \mathcal{Z}^n \to \ddot{\mathcal{Y}}$ through a Gradient Reversal Layer (GRL) [14] (Fig. 2). The classifier attempts to infer the ordinal label from z^n, while the GRL inverts feedback during backpropagation, forcing the quantizer to suppress order-sensitive signals. The classifier is supervised by SOR, with $\ell_{\mathrm{C}}^{\mathrm{n}}(z^n, \ddot{y}) = \mathrm{SOR}(\mathsf{n}(\mathrm{GRL}(z^n)), \ddot{y})$, thereby adversarially discouraging the non-ordinal codebook from encoding rank information.

Quantizer Losses. The ordinal and non-ordinal codebooks are optimized with standard vector-quantization losses $\ell_{\mathrm{Q}}^{\mathrm{o}}(z_1, z_2)$ and $\ell_{\mathrm{Q}}^{\mathrm{n}}(z_1, z_2)$ [50], both following $\|\mathrm{sg}[z_1] - z_2\|_2^2 + \|z_1 - \mathrm{sg}[z_2]\|_2^2$, where $\mathrm{sg}[\cdot]$ denotes the stop-gradient. The ordinal loss uses $(z_1, z_2) = (\tilde{z}, z^o)$, while the non-ordinal loss uses $(z_1, z_2) = (z, z^n)$. To discourage correlation and promote disentanglement, we additionally minimize the cosine similarity between the two latent representations, $\ell_{\mathrm{sim}}{=}\eta(z^{\mathrm{o}}, z^{\mathrm{n}})$

Decoder and Discriminator. The decoder $\mathsf{g} : \mathcal{Z}^o \times \mathcal{Z}^n \to \mathcal{X}$ reconstructs the image from the concatenated quantized representations as $\hat{x}{=}\mathsf{g}(z^o, z^n)$ and is optimized to minimize pixel reconstruction error and perceptual distance with $\ell_{\mathrm{rec}}(x, \hat{x}){=}\|x{-}\hat{x}\|_2^2 + \mathrm{LPIPS}(x, \hat{x})$ [58].

The discriminator $\mathsf{d} : \mathcal{X} \to \{0, 1\} \times \ddot{\mathcal{Y}}$ adopts a PatchGAN architecture [24] with shared convolutional layers and two task-specific heads. The real/fake head d_{rf} is optimized with L_2 objective, $\ell_{\mathsf{d}_{\mathrm{rf}}}(x, \hat{x}){=}\big\|\mathsf{d}_{\mathrm{rf}}(x) - 1\big\|_2^2 + \big\|\mathsf{d}_{\mathrm{rf}}(\hat{x})\big\|_2^2$ while the classification head $\mathsf{d}_{\mathrm{clf}}$ is supervised with SOR, $\ell_{\mathsf{d}_{\mathrm{clf}}}(x, \ddot{y}){=}\mathrm{SOR}(\mathsf{d}_{\mathrm{clf}}(x), \ddot{y})$. Together, these objectives encourage perceptually realistic reconstructions that preserve ordinal semantics.

4.2 DisMix Training

The generator, comprising encoder $\mathsf{e}(.)$, codebooks $\mathbf{M}^o, \mathbf{M}^n$, and decoder $\mathsf{g}(.)$, is trained adversarially against the discriminator $\mathsf{d}(.)$ via:

$$\min_{\mathsf{e}, \mathbf{M}^o, \mathbf{M}^n, \mathsf{g}} \max_{\mathsf{d}} \; \mathbb{E}_x\big[\ell_{(\mathsf{e}, \mathbf{M}^o, \mathbf{M}^n, \mathsf{g})} - \ell_{\mathsf{d}}\big] \tag{1}$$

The generator loss $\ell_{(\mathsf{e}, \mathbf{M}^o, \mathbf{M}^n, \mathsf{g})}$ combines quantization ($\ell_{\mathrm{Q}}^{\mathrm{o}}, \ell_{\mathrm{Q}}^{\mathrm{n}}$), classification ($\ell_{\mathrm{C}}^{\mathrm{o}}, \ell_{\mathrm{C}}^{\mathrm{n}}$), reconstruction ($\ell_{\mathrm{rec}}$), and similarity ($\ell_{\mathrm{sim}}$) terms, while the discriminator loss ℓ_{d} combines realism ($\ell_{\mathsf{d}_{\mathrm{rf}}}$) and ordinal classification ($\ell_{\mathsf{d}_{\mathrm{clf}}}$) objectives. All terms are weighted by learnable coefficients that are jointly optimized with the model under a unit-sum constraint, eliminating the need for manual hyperparameter tuning.

4.3 Order Aware Mixup with Disentangled Latents

Once trained, DisMix is frozen and used to generate mixed samples during downstream ordinal classifier training. Given two images x_i, x_j from adjacent ranks $\bar{y}_i, \bar{y}_j$ determined by majority grade (i.e., $\bar{y}_i{=}\mathrm{majority}(\{y_{i,a}\}_{a \in \mathcal{A}})$ and similarly for $\bar{y}_j$), with corresponding ordinal and non-ordinal codes (z_i^o, z_i^n) and (z_j^o, z_j^n) extracted from quantizers, and a mixing ratio $\lambda \in (0, 1)$, Tab. 1 defines four mixing policies.

Ordinal Mix interpolates ordinal factors while keeping non-ordinal features fixed. Formally, $\tilde{z}^o{=}\lambda z_i^o{+}(1{-}\lambda)z_j^o, \quad \tilde{z}^n{\in}\{z_i^n, z_j^n\}$. Then we obtain the augmented sample by, $\hat{x}{=}\mathsf{g}(\tilde{z}^o, \tilde{z}^n)$, and its label $\tilde{y}{=}\lambda\bar{y}_i{+}(1{-}\lambda)\bar{y}_j$. *Non-Ordinal Mix* instead interpolates order-invariant features and takes the majority grade of the sample that provides the ordinal code. *Order Swap* swaps the ordinal and non-ordinal features between samples. $(\tilde{z}^{\mathrm{o}}, \tilde{z}^{\mathrm{n}}) \in \{(z_j^o, z_i^n), (z_i^o, z_j^n)\}, \quad \hat{x} = \mathsf{g}(\tilde{z}^{\mathrm{o}}, \tilde{z}^{\mathrm{n}})$, and $\tilde{y}$ is the rank associated with the chosen $\tilde{z}^{\mathrm{o}}$.

Generate & Mix generates an adjacent-rank variant of an anchor and applies Ordinal Mix. We first select either x_i or x_j as an anchor and obtain an adjacent rank from its majority grade. By

Table 2: Comparing DisMix against six image mixup baselines—Mixup, CutMix, PuzzleMix, CoMix, GuidedMix, DiffuseMix across six ordinal classification methods—POE, GOL, MWR, RnC, ORD2SEQ, CLOC using four datasets—IDRID, KOA, Chaoyang and Gleason.

	Accuracy ↑						Mean Absolute Error (MAE) ↓					
	POE	GOL	MWR	RNC	ORD2SEQ	CLOC	POE	GOL	MWR	RNC	ORD2SEQ	CLOC
IDRID												
Mixup	57.93 ± 4.05	55.34 ± 1.68	39.16 ± 1.13	54.04 ± 3.12	63.43 ± 0.56	64.08 ± 0.97	0.72 ± 0.06	0.52 ± 0.03	0.78 ± 0.02	0.79 ± 0.07	0.64 ± 0.04	1.11 ± 0.20
CutMix	58.27 ± 0.06	56.31 ± 2.57	40.76 ± 1.70	55.94 ± 2.06	65.37 ± 0.56	64.69 ± 1.53	0.64 ± 0.02	0.56 ± 0.05	0.78 ± 0.05	0.75 ± 0.05	0.62 ± 0.03	0.58 ± 0.01
PuzzleMix	59.80 ± 3.75	55.99 ± 1.48	39.16 ± 1.49	56.05 ± 3.34	66.00 ± 1.00	64.72 ± 2.97	**0.59 ± 0.00**	0.53 ± 0.02	0.81 ± 0.03	0.78 ± 0.04	0.62 ± 0.01	0.58 ± 0.08
CoMix	58.90 ± 2.14	58.58 ± 3.12	40.78 ± 1.68	54.04 ± 2.97	65.95 ± 1.74	64.39 ± 0.56	0.64 ± 0.00	0.50 ± 0.02	0.81 ± 0.03	0.82 ± 0.07	0.62 ± 0.03	0.58 ± 0.03
GuidedMix	56.33 ± 0.06	70.22 ± 2.02	38.83 ± 1.69	53.07 ± 2.02	64.73 ± 1.48	64.07 ± 2.57	0.73 ± 0.01	0.33 ± 0.02	0.84 ± 0.04	0.85 ± 0.04	0.68 ± 0.03	0.61 ± 0.05
DiffuseMix	58.27 ± 0.06	66.64 ± 2.03	40.77 ± 0.96	52.08 ± 1.14	65.70 ± 1.48	63.75 ± 1.12	0.69 ± 0.01	0.35 ± 0.03	0.83 ± 0.05	0.95 ± 0.05	0.62 ± 0.02	0.66 ± 0.02
DisMix	**60.53 ± 1.99**	**73.46 ± 3.41**	**41.95 ± 0.35**	**56.31 ± 0.98**	**66.34 ± 1.12**	**66.02 ± 0.98**	0.66 ± 0.01	**0.29 ± 0.06**	**0.77 ± 0.02**	**0.73 ± 0.03**	**0.58 ± 0.02**	**0.57 ± 0.02**
KOA												
Mixup	68.40 ± 0.26	76.89 ± 0.89	29.57 ± 25.23	63.96 ± 0.55	67.92 ± 0.26	67.16 ± 0.84	0.41 ± 0.01	0.23 ± 0.01	0.61 ± 0.02	**0.49 ± 0.00**	0.43 ± 0.01	0.42 ± 0.01
CutMix	68.20 ± 0.26	76.75 ± 0.26	44.40 ± 0.55	62.90 ± 0.24	69.03 ± 0.57	67.59 ± 0.24	**0.38 ± 0.00**	0.23 ± 0.00	0.61 ± 0.02	0.51 ± 0.00	0.43 ± 0.02	0.42 ± 0.00
PuzzleMix	68.37 ± 0.57	77.17 ± 0.10	43.56 ± 0.97	62.90 ± 0.23	69.18 ± 0.51	67.02 ± 0.21	0.40 ± 0.01	0.23 ± 0.00	0.63 ± 0.07	0.51 ± 0.01	0.41 ± 0.02	0.42 ± 0.01
CoMix	68.93 ± 0.75	77.21 ± 0.53	44.24 ± 0.49	**63.58 ± 0.90**	68.64 ± 0.24	67.79 ± 0.53	0.39 ± 0.03	0.23 ± 0.01	0.62 ± 0.04	0.49 ± 0.02	0.41 ± 0.01	**0.40 ± 0.02**
GuidedMix	68.07 ± 1.14	76.91 ± 0.15	44.40 ± 0.23	63.36 ± 0.35	68.92 ± 0.31	67.57 ± 0.43	0.41 ± 0.01	0.23 ± 0.00	0.61 ± 0.02	0.50 ± 0.01	0.43 ± 0.03	0.42 ± 0.01
DiffuseMix	68.40 ± 0.44	77.40 ± 0.46	43.98 ± 0.18	63.54 ± 0.43	68.68 ± 0.31	67.64 ± 0.98	0.39 ± 0.01	0.23 ± 0.00	0.60 ± 0.03	0.49 ± 0.01	0.43 ± 0.02	0.41 ± 0.01
DisMix	**69.80 ± 1.04**	**77.72 ± 0.16**	**44.53 ± 0.25**	63.28 ± 0.76	**69.41 ± 0.45**	**68.41 ± 0.36**	0.40 ± 0.01	**0.22 ± 0.00**	**0.57 ± 0.05**	0.49 ± 0.02	**0.40 ± 0.02**	0.41 ± 0.01
CHAOYANG												
Mixup	82.67 ± 0.65	88.29 ± 0.40	57.73 ± 1.20	82.53 ± 0.51	80.38 ± 2.66	85.36 ± 0.13	0.24 ± 0.01	0.13 ± 0.00	0.46 ± 0.01	0.25 ± 0.01	0.26 ± 0.02	0.21 ± 0.00
CutMix	82.40 ± 0.30	88.21 ± 0.27	56.64 ± 0.51	82.12 ± 0.64	53.06 ± 3.56	84.91 ± 0.10	0.24 ± 0.00	0.13 ± 0.00	0.48 ± 0.01	0.26 ± 0.01	0.86 ± 0.04	0.22 ± 0.00
PuzzleMix	83.17 ± 0.35	87.95 ± 0.35	57.71 ± 0.59	81.72 ± 0.57	54.81 ± 2.02	85.47 ± 0.49	0.23 ± 0.00	0.13 ± 0.00	0.47 ± 0.02	0.27 ± 0.01	0.84 ± 0.02	0.21 ± 0.01
CoMix	83.43 ± 0.25	88.10 ± 0.41	57.52 ± 0.71	81.84 ± 0.17	55.74 ± 3.51	85.18 ± 1.97	0.23 ± 0.01	0.13 ± 0.01	0.47 ± 0.00	0.26 ± 0.00	0.89 ± 0.02	0.21 ± 0.03
GuidedMix	83.13 ± 0.35	87.72 ± 0.38	56.80 ± 0.33	81.64 ± 0.10	57.95 ± 1.90	**85.72 ± 0.33**	0.23 ± 0.01	0.13 ± 0.00	0.47 ± 0.01	0.27 ± 0.00	0.83 ± 0.01	**0.21 ± 0.00**
DiffuseMix	82.23 ± 0.06	87.56 ± 0.14	57.57 ± 0.45	80.87 ± 0.35	83.67 ± 0.51	85.39 ± 0.48	0.24 ± 0.01	0.13 ± 0.00	**0.46 ± 0.00**	0.28 ± 0.02	**0.25 ± 0.01**	0.21 ± 0.01
DisMix	**84.10 ± 0.52**	**88.58 ± 0.26**	**58.06 ± 0.19**	**83.30 ± 0.45**	**84.27 ± 0.17**	84.88 ± 0.66	**0.22 ± 0.01**	**0.12 ± 0.00**	0.47 ± 0.00	**0.24 ± 0.01**	0.26 ± 0.00	0.22 ± 0.01
GLEASON												
Mixup	77.40 ± 0.62	93.96 ± 0.60	79.68 ± 6.27	87.78 ± 2.07	92.46 ± 0.57	88.93 ± 1.92	0.32 ± 0.01	0.05 ± 0.00	0.58 ± 0.29	0.17 ± 0.02	0.24 ± 0.06	0.15 ± 0.02
CutMix	83.30 ± 4.15	94.88 ± 0.05	**84.19 ± 2.96**	88.99 ± 1.01	92.65 ± 0.31	89.09 ± 0.65	0.33 ± 0.05	0.05 ± 0.00	**0.21 ± 0.05**	0.16 ± 0.02	0.28 ± 0.02	0.16 ± 0.02
PuzzleMix	81.43 ± 1.01	95.21 ± 0.39	71.60 ± 5.95	88.31 ± 0.45	92.23 ± 0.28	88.95 ± 1.73	0.33 ± 0.04	0.05 ± 0.00	0.22 ± 0.01	0.17 ± 0.01	**0.24 ± 0.02**	0.17 ± 0.03
CoMix	83.47 ± 1.67	94.77 ± 0.43	67.78 ± 1.43	88.88 ± 1.38	92.44 ± 0.28	88.45 ± 0.82	0.31 ± 0.05	0.06 ± 0.01	0.65 ± 0.10	0.16 ± 0.00	0.26 ± 0.02	0.17 ± 0.01
GuidedMix	84.10 ± 2.44	95.05 ± 0.45	78.57 ± 5.04	88.87 ± 0.45	92.25 ± 0.29	88.64 ± 0.77	0.34 ± 0.01	0.05 ± 0.01	0.25 ± 0.04	0.17 ± 0.01	0.26 ± 0.05	0.17 ± 0.01
DiffuseMix	85.63 ± 3.25	**95.28 ± 0.36**	77.94 ± 11.6	88.81 ± 0.81	92.74 ± 0.72	89.79 ± 1.78	0.33 ± 0.04	0.05 ± 0.00	0.46 ± 0.15	0.16 ± 0.02	0.28 ± 0.03	0.15 ± 0.02
DisMix	**86.30 ± 3.84**	94.77 ± 0.23	76.89 ± 8.60	**89.22 ± 0.42**	**93.26 ± 1.15**	**90.77 ± 0.93**	**0.30 ± 0.03**	**0.04 ± 0.01**	0.23 ± 0.01	**0.15 ± 0.01**	0.27 ± 0.06	**0.14 ± 0.02**

conditioning the anchor's ordinal code on the adjacent rank, we obtain $z^{o,\text{cnd}}$ from the ordinal quantizer. The non-ordinal code $z^{n,anc}$ of the anchor is extracted via the non-ordinal quantizer to produce $x'=\mathsf{g}\big(z^{o,\text{cnd}},\ z^{n,\text{anc}}\big)$, with y' corresponding to the conditioned rank. Ordinal Mix is then applied between the anchor and x'.

Non-Ordinal Mix and Order Swap are label-preserving policies, retaining the rank of the source sample providing the ordinal code. Ordinal Mix and Generate & Mix are label-interpolating policies, assigning soft labels between the two source ranks.

5 Experiments

Datasets And Preparation. The **Knee Osteoarthritis (KOA)** dataset [8] includes knee X-ray images (6,604 train/1,656 test) graded on the Kellgren–Lawrence (KL) scale (0–4). The **Indian Diabetic Retinopathy (IDRID)** [43] contains retinal fundus images (413/103) graded on the DR scale (0–4). The **Chaoyang** dataset [61] has colorectal histopathology images (4,021/2,139) graded by three pathologists, with $\sim$40% disagreement in training and consensus test labels. The **Crowd Gleason** dataset [39] contains prostate histopathology patches (16,151/2,926) graded by seven residents (mean κ=0.54) on the Gleason scale (NC, G3–G5), with consensus test labels. More in App. B.

Images are resized to 256×256, normalized to $[-1, 1]$, and augmented with random flips and rotations $< 20^\circ$; with histogram equalization for KOA. IDRID and KOA simulate five labelers following [42], randomly flipping 60% grade 1→2 and 20% grade 2→1, resulting in an overall 30% grading variability in the dataset. Chaoyang and Gleason contain real grading variability. Each training image was assigned an ordinal soft label for DisMix and a generic soft label for downstream classifiers (POE, GOL, etc.) during training, while test labels remained fixed.

DisMix Setup. We adopt the encoder, decoder, and quantizer architectures of [45] with pretrained `vq-f8` weights. The non-ordinal and ordinal codebooks use 16,384 and 512, 4-dim entries, respectively, following pretrained compatibility and the smaller ordinal subspace [59]. The discriminator implements [45], with a linear classifier attached before the real/fake convolution for rank prediction. DisMix is trained for 400 epochs using Adam (1×10^{-5}, batch size 11), freezing encoder downsampling blocks. The checkpoint that maximizes mixing accuracy, measuring agreement between discriminator-predicted ranks and ranks expected from random latent interpolation, is selected for downstream tasks.

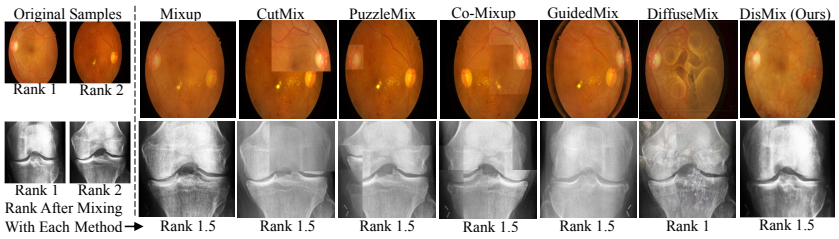


Figure 3: Mixing a severe (rank 2) sample into a mild (rank 1) sample from IDRID (top) and KOA (bottom). DisMix produces plausible ordinal transitions from mild samples—such as darker lesion regions in IDRID and reduced bone gap in KOA—while preserving non-ordinal attributes like eye orientation and bone structure. In contrast, other mixup often generate unrealistic structures, including multiple retinal optic discs, disoriented bone gaps, and imaging artifacts.

Downstream Ordinal Model Setup. Ordinal baselines use their original implementation, extending loss functions to support generic soft labels where necessary. Models are trained for 200 epochs using Adam (batch size 64), with 15% of the training set reserved for validation and the best validation checkpoint used for testing.

Baselines. We evaluate on six ordinal classification methods (POE[37], GOL[36], MWR[47], RnC[56], ORD2SEQ[53], and CLOC[42]) across four datasets, comparing DisMix with six state-of-the-art (SOTA) mixup methods: Mixup[57], CutMix[55], PuzzleMix[32], CoMix[31], GuidedMix[27], and DiffuseMix[23].

Evaluation Criteria. We report the mean and standard deviation of Accuracy and MAE over three runs, using a mixing probability of 0.5 unless otherwise specified. DisMix is evaluated by comparing the gains of ordinal algorithms when paired with DisMix versus alternative mixup methods.

5.1 Comparison with SOTA.

Table 2 shows that, DisMix achieves the best mean accuracy and MAE in 20 and 15 of 24 settings, each defined by applying DisMix with an ordinal algorithm on a dataset. Compared with the strongest baseline, these wins significantly improves accuracy (one-sided Wilcoxon signed-rank test, $p = 0.0075$) while typically reducing MAE by 4–6%, whereas other mixups often trade accuracy for MAE. Surpassing diffusion-based methods like DiffuseMix underscores the benefit of explicitly modeling ordinal features.

Figure 3 compares mixed samples generated by different methods on IDRID and KOA. Existing methods often produce unrealistic outputs, including duplicate optic discs and abnormal anatomical structures. In contrast, DisMix generates plausible severity transitions, such as darker lesions and narrower joint spaces, while preserving non-ordinal attributes including eye orientation and bone structure, demonstrating the benefit of mixing in disentangled latent spaces.

5.2 Boundary Error Reduction.

Beyond accuracy and MAE, we evaluate boundary-level classification errors using CLOC across datasets. Figure 4 shows that DisMix consistently achieves the lowest boundary errors, notably reducing IDRID C2↔C3 by 9%, Chaoyang C1↔C2 by 4%, Gleason C2↔C3 6% over to the best mixup baseline.

5.3 Performance with Scarce Data.

We restrict the training set to 10% of samples per class (minimum 10 images) for both DisMix and CLOC. Table 3 shows that DisMix outperforms baseline mixup methods. Notably, on IDRID, DisMix improves accuracy by $\approx$1.5% over the best baseline with only 60 original training images, demonstrating strong data efficiency and robustness in low-data regimes. More in App. C.

Table 3: Performance under data scarcity, training set limited to 10% of samples/class ($\geq$10 images/class).

	Accuracy ↑	MAE ↓
IDRID		
Mixup	43.36 ± 1.49	1.06 ± 0.04
GuidedMix	46.09 ± 1.77	1.00 ± 0.09
DiffuseMix	45.26 ± 2.84	1.13 ± 0.08
DisMix	**47.91 ± 2.82**	**0.99 ± 0.09**
GLEASON		
Mixup	84.82 ± 0.95	0.21 ± 0.02
GuidedMix	84.22 ± 0.22	0.23 ± 0.01
DiffuseMix	84.67 ± 1.46	0.23 ± 0.01
DisMix	**85.38 ± 0.76**	**0.21 ± 0.01**

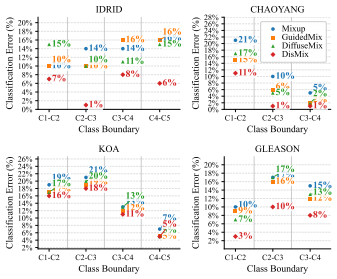


Figure 4: Per-boundary classification error rates of DisMix vs. baseline mixup methods.

Table 4: DisMix's performance heatmaps under grading variability (30%, 50%, 70%) and mix probabilities (0.3, 0.5, 0.7).

Vari. Level	Accuracy ↑			MAE ↓		
IDRID						
70%	56.99	58.89	55.33	0.67	0.65	0.72
50%	65.37	62.78	67.98	0.57	0.61	0.52
30%	66.92	66.02	64.73	0.57	0.57	0.64
KOA						
70%	64.46	65.57	64.85	0.47	0.45	0.46
50%	67.93	68.19	68.55	0.42	0.41	0.41
30%	68.03	68.41	68.09	0.42	0.41	0.41
Mix Prob.	0.3	0.5	0.7	0.3	0.5	0.7

5.4 Robustness to Grading Variability.

Following [42], we increase the boundary grade variability in IDRID and KOA by flipping 70/40% and 90/60% of grade 1→2 / 2→1 (yielding two settings with 50% and 70% overall grading variability for each dataset), and vary the mixup probability to 0.3, 0.5, and 0.7 with CLOC. Tab. 4 shows DisMix performs best overall with a mix probability of 0.5, whereas 0.3 is preferable under mild grading variability (30%) by providing sufficient data diversification. Under severe variability (70%), accuracy drops and MAE increases, especially at 0.7, indicating excessive mixing amplifies boundary-level variability. More in App. C.

6 Ablation Studies and Discussion

Quality of Disentanglement. The rightmost panel of Fig.1 shows the disentanglement on KOA, with the ordinal codebook exhibiting a rank-aligned structure and the non-ordinal codebook no apparent rank organization. We quantify this by training linear probes on the frozen latent features (z^o and z^n) and reporting test-set rank prediction accuracy. As shown in Tab.5, the probe achieves 65.04% on z^o vs. 19.65% on z^n, confirming that ordinal information is concentrated in z^o, while z^n remains at chance for five-class KOA. An ablation study further shows that removing the GRL or ℓ_{sim} increases rank leakage into z^n, whereas removing the dual codebooks collapses accuracy to 33%, validating the proposed design.

Figure 5 illustrates that the disentanglement captures clinically meaningful factors of knee osteoarthritis grading. Joint-space narrowing, a key severity indicator, is encoded in the ordinal latent, whereas the surgical nail, unrelated to grading, is captured in the non-ordinal latent [3, 16]. Accordingly, modifying the non-ordinal latent alters the nail (rows 2 and 4), while modifying the ordinal latent changes the joint space (row 3), demonstrating a controllable and semantically meaningful decomposition.

Analysis of DisMix Components and Policies. To assess the contribution of each component, we evaluate the proposed mixup policies individually. Ordinal Mix, Non-Ordinal Mix, Generate & Mix, and Order Swap isolate the ordinal branch, non-ordinal branch, conditional rank generation, and learned feature exchange, respectively. Table 6 shows, each policy is competitive against each other, but remains weaker than their combined use in Tab.2, highlighting their complementary roles. Tab.12 (App. C) finds that Generate & Mix achieves higher accuracy in severe data scarcity (10% images/class) and extreme grading variability (70%) likely due to its two-stage process of generating a clean adjacent rank before mixing, highlighting the benefit of ordinal-conditioned synthesis. Tab.11 (App. C) further group policies based on label-preserving and label-interpolating and show that label-preserving policies slightly outperform interpolating policies when used individually.

Without Grading Variability. Although DisMix was evaluated primarily with grading variability, we also evaluate without grading variability (i.e. a single label per image). Table 7 shows that DisMix achieves the best accuracy and MAE on both IDRID and KOA, indicating that DisMix remains effective even in the absence of grading variability.

Table 5: Probe classifier performance with DisMix variants.

	Accuracy ↑		MAE ↓	
	on z^o	on z^n	on z^o	on z^n
DisMix	65.04	19.65	0.45	1.42
Without GRL	62.13	22.95	0.48	1.36
Without ℓ_{orth}	59.22	24.35	0.59	1.38
One codebook	33.01		1.72	

Table 6: Performance vs. individual mixing policies with CLOC.

	Accuracy ↑	MAE ↓
	IDRID	
Ordinal Mix	65.09 ± 1.73	0.56 ± 0.02
Generate & Mix	67.01 ± 2.56	0.55 ± 0.04
Non-Ordinal Mix	64.74 ± 1.17	0.60 ± 0.01
Order Swap	66.69 ± 2.96	0.54 ± 0.06
	GLEASON	
Ordinal Mix	88.96 ± 1.31	0.17 ± 0.01
Generate & Mix	90.34 ± 1.43	0.13 ± 0.02
Non-Ordinal Mix	89.62 ± 1.19	0.16 ± 0.02
Order Swap	90.39 ± 0.97	0.14 ± 0.02

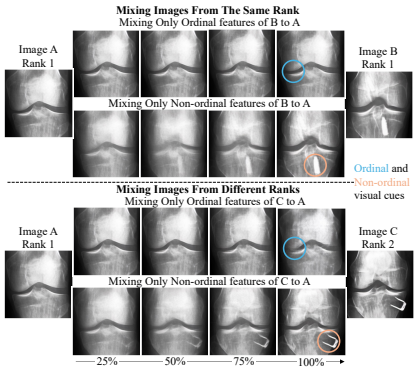


Figure 5: Mixing either ordinal or non-ordinal features into reference image (A) from same-rank (B) and severe-rank (C) KOA samples at varying ratios. Ordinal mixing from (B) preserves the joint space, while (C) narrows it; non-ordinal mixing mainly alters appearance (e.g., nails).

Table 7: DisMix vs. mixup baselines with CLOC with no grading variability.

	Accuracy ↑	MAE ↓
	IDRID	
Mixup	65.37 ± 2.25	0.58 ± 0.02
CutMix	64.39 ± 0.56	0.59 ± 0.06
PuzzleMix	65.69 ± 0.56	0.56 ± 0.00
CoMix	65.04 ± 0.00	0.56 ± 0.02
GuidedMix	64.72 ± 1.48	0.56 ± 0.01
DiffuseMix	65.69 ± 2.02	0.58 ± 0.06
DisMix	**66.01 ± 0.98**	**0.55 ± 0.03**
	KOA	
Mixup	68.55 ± 0.37	0.39 ± 0.01
CutMix	68.97 ± 0.82	0.38 ± 0.02
PuzzleMix	68.45 ± 0.35	0.38 ± 0.02
CoMix	68.37 ± 0.69	0.39 ± 0.01
GuidedMix	68.49 ± 0.09	0.39 ± 0.01
DiffuseMix	68.35 ± 0.38	0.39 ± 0.02
DisMix	**69.53 ± 0.47**	**0.38 ± 0.01**

Scope of Generations. Although Fig. 3 and 5 show that DisMix produces visually plausible ordinal and non-ordinal variations, the generated images are intended for data augmentation rather than direct clinical interpretation, as the generator prioritizes preserving ordinal semantics over high-fidelity image synthesis. Furthermore, the interpolated ranks are used solely for soft supervision and should not be interpreted as new clinically defined grades.

Limitations. We measure mixup augmentation overhead as the percentage increase in downstream training time relative to vanilla training (without mixup) on IDRID using GOL. Figure 6 shows that DisMix offers a good accuracy–efficiency trade-off, outperforming mixup baselines while incurring lower overhead than DiffuseMix and Co-Mixup. This can be further mitigated by generating and caching mixed samples offline before downstream training. Furthermore, training DisMix incurs a one-time, per-dataset overhead to learn the disentangled representation, which is amortized across all downstream classifiers (see App. C). Future work will explore foundation models for few-shot ordinal disentanglement to reduce per-dataset training.

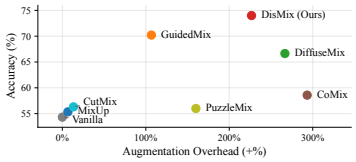


Figure 6: Accuracy vs. augmentation overhead of DisMix vs. mixup baselines on IDRID.

7 Conclusion

We presented DisMix, a latent mixup method for ordinal data motivated by the observation that conventional image mixup entangles order-relevant and order-irrelevant factors, thereby weakening the ordinal signal. DisMix disentangles ordinal and non-ordinal features into separate latent subspaces via a dual-codebook VQ-VAE, enabling order-aware mixing within dedicated subspaces. Extensive experiments show that DisMix achieves strongest aggregate performance among six SOTA mixup baselines across six ordinal classifiers, improving both accuracy and MAE. Our results underscore the importance of order-aware mixup for ordinal data highlighting disentanglement-based methods as a promising direction for robust ordinal learning in medical imaging.

References

[1] B. Bae, J. Lee, S. T. Kong, J. Sung, and K.-H. Jung. Manifold ordinal-mixup for ordered classes in tw3-based bone age assessment. In A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 662–670, Cham, 2020. Springer International Publishing.

[2] G. Baykal, M. Kandemir, and G. Unal. Disentanglement with factor quantized variational autoencoders. *Neurocomputing*, 661:131968, 2026.

[3] H. J. Braun and G. E. Gold. Diagnosis of osteoarthritis: Imaging. *Bone*, 51(2):278–288, Aug. 2012.

[4] Z. Cao, E. Chen, Y. Huang, S. Shen, and Z. Huang. Learning from Crowds with Annotation Reliability. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Taipei Taiwan, July 2023. ACM.

[5] G. Carneiro. *Machine Learning with Noisy Labels: Definitions, Theory, Techniques and Solutions*. Elsevier, 2024.

[6] G. Chen, S. Zhang, D. Lin, H. Huang, and P. A. Heng. Learning to aggregate ordinal labels by maximizing separating width. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 787–796. PMLR, 06–11 Aug 2017.

[7] H. Chen, Z. Wang, X. Li, X. Sun, F. Chen, J. Liu, J. Wang, B. Raj, Z. Liu, and E. Barsoum. Softvq-vae: Efficient 1-dimensional continuous tokenizer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28358–28370, 2025.

[8] P. Chen, L. Gao, X. Shi, K. Allen, and Y. Lin. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, 75:84–92, 2019.

[9] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[10] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

[11] R. Diaz and A. Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4738–4747, 2019.

[12] Y. Dong, F.-Y. Su, and J.-H. Chiang. Sgd-mix: Enhancing domain-specific image classification with label-preserving data augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7051–7061, March 2026.

[13] A. Galdran, G. Carneiro, and M. A. González Ballester. Balanced-mixup for highly imbalanced medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 323–333. Springer, 2021.

[14] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[15] H. W. Goh, U. Tkachenko, and J. Mueller. CROWDLAB: Supervised learning to infer consensus labels and quality scores for data with multiple annotators. In *NeurIPS Human in the Loop Learning Workshop*, 2022.

[16] D. Hayashi, F. W. Roemer, M. Jarraya, and A. Guermazi. Imaging in osteoarthritis. *Radiologic Clinics of North America*, 55(5):1085–1102, Sept. 2017.

[17] M. Herde, D. Huseljic, and B. Sick. Multi-annotator deep learning: A probabilistic framework for classification. *Transactions on Machine Learning Research*, 2024.

[18] M. Herde, L. Lührs, D. Huseljic, and B. Sick. Annot-Mix: Learning with noisy class labels from multiple annotators via a mixup extension. In *European Conference on Artificial Intelligence*, 2024.

[19] X. Hou, Y. Li, and S. Wang. Disentangled representation for age-invariant face recognition: A mutual information minimization perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3692–3701, 2021.

[20] K. Hsu, W. Dorrell, J. C. R. Whittington, J. Wu, and C. Finn. Disentanglement via latent quantization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[21] K. Hsu, J. I. Hamid, K. Burns, C. Finn, and J. Wu. Tripod: Three complementary inductive biases for disentangled representation learning. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19101–19122. PMLR, 21–27 Jul 2024.

[22] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[23] K. Islam, M. Z. Zaheer, A. Mahmood, and K. Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27621–27630, 2024.

[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[25] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12341–12351, 2021.

[26] Y. Jun, J. Park, K. Choo, T. E. Choi, and S. J. Hwang. Disentangling disentangled representations: Towards improved latent units via diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3559–3569. IEEE, 2025.

[27] M. Kang and S. Kim. Guidedmixup: an efficient mixup strategy guided by saliency maps. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 1096–1104, 2023.

[28] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[29] V. Khrulkov, L. Mirvakhabova, I. Oseledets, and A. Babenko. Disentangled representations from non-disentangled models. *arXiv preprint arXiv:2102.06204*, 2021.

[30] H. Kim and A. Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018.

[31] J. Kim, W. Choo, H. Jeong, and H. O. Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. In *International Conference on Learning Representations*, 2021.

[32] J.-H. Kim, W. Choo, and H. O. Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International conference on machine learning*, pages 5275–5285. PMLR, 2020.

[33] K. Kobayashi, R. Hataya, Y. Kurose, M. Miyake, M. Takahashi, A. Nakagawa, T. Harada, and R. Hamamoto. Decomposing normal and abnormal features of medical images for content-based image retrieval of glioma imaging. *Medical Image Analysis*, 74:102227, 2021.

[34] G. Kwon and J. C. Ye. Diffusion-based image translation using disentangled style and content representation. In *The Eleventh International Conference on Learning Representations*, 2023.

[35] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato. Fader networks: Manipulating images by sliding attributes. *Advances in neural information processing systems*, 30, 2017.

[36] S.-H. Lee, N. H. Shin, and C.-S. Kim. Geometric order learning for rank estimation. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[37] W. Li, X. Huang, J. Lu, J. Feng, and J. Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13896–13905, 2021.

[38] Y. Liu, A. W. K. Kong, and C. K. Goh. A constrained deep neural network for ordinal regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 831–839, 2018.

[39] M. López-Pérez, A. Morquecho, A. Schmidt, F. Pérez-Bueno, A. Martín-Castro, J. Mateos, and R. Molina. The crowdgleason dataset: Learning the gleason grade from crowds and experts. *Computer Methods and Programs in Biomedicine*, 257:108472, 2024.

[40] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[41] J. Noh, H. Park, J. Lee, and B. Ham. Rankmixup: Ranking-based mixup training for network calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1358–1368, 2023.

[42] D. Pitawela, G. Carneiro, and H.-T. Chen. CLOC: Contrastive learning for ordinal classification with multi-margin n-pair loss. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15538–15548, 2025.

[43] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data*, 3(3), 2018.

[44] H. Qin, X. Jin, H. Zhu, H. Liao, M. A. El-Yacoubi, and X. Gao. Sumix: Mixup with semantic and uncertain information. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024.

[45] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[46] Y. Shen and B. Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1532–1540, 2021.

[47] N.-H. Shin, S.-H. Lee, and C.-S. Kim. Moving window regression: A novel approach to ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18760–18769, 2022.

[48] V. B. Sinha, S. Rao, and V. N. Balasubramanian. Fast Dawid-Skene: A fast vote aggregation scheme for sentiment classification. In *KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, 2018.

[49] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017.

[50] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[51] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[52] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019.

[53] J. Wang, Y. Cheng, J. Chen, T. Chen, D. Chen, and J. Wu. Ord2seq: Regarding ordinal regression as label sequence prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5875, 2023.

[54] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9677–9696, 2024.

[55] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[56] K. Zha, P. Cao, J. Son, Y. Yang, and D. Katabi. Rank-n-contrast: Learning continuous representations for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[57] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[59] J. Zheng, Y. Yao, B. Han, D. Wang, and T. Liu. Enhancing contrastive learning for ordinal regression via ordinal content preserved data augmentation. In *The Twelfth International Conference on Learning Representations*, 2024.

[60] D. Zhou, Q. Liu, J. Platt, and C. Meek. Aggregating ordinal labels from crowds by minimax conditional entropy. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 262–270, Bejing, China, 22–24 Jun 2014. PMLR.

[61] C. Zhu, W. Chen, T. Peng, Y. Wang, and M. Jin. Hard sample aware noise robust learning for histopathology image classification. *IEEE Transactions on Medical Imaging*, 41(4):881–894, 2022.

[62] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.

[63] K. Zou, S. Faisan, F. Heitz, and S. Valette. Disentangling high-level factors and their features with conditional vector quantized vaes. *Pattern Recognition Letters*, 172:172–180, 2023.

Appendix

A Ordinal Soft Label

This section elaborates on the ordinal soft-label creation process with a fail-safe mechanism and examples.

Each sample x_i, graded by more than one annotators with scalar ranks $\{y_{i,a}\}_{a\in\mathcal{A}}$, is assigned a single ordinal soft label. Firstly, each label $y_{i,a}$ is converted into a progressive vector $\mathbf{p}_{i,a}^{(k)} = \mathbb{I}(k < \tilde{y}_{i,a})$ for $k \in [0, C-2]$. Secondly, taking the element-wise average across annotators yields $\mathbf{s}_i^{(k)}=\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} p_{i,a}^{(k)}$, forming $\mathbf{s}_i=[s_i^{(0)}, \ldots, s_i^{(C-2)}]$. Lastly, to ensure a non-increasing progression, an optional monotonicity constraint is applied to $\mathbf{s}_i$, ensuring $s_i^{(k+1)} \leq s_i^{(k)}$ for all k. The resulting target $\ddot{y}_i$ reflects that belonging to a higher rank (e.g., 4) implies inclusion in all preceding ranks. For instance, if C=5, $|\mathcal{A}|$=4, and annotations,

- $\{y_{i,a}\}_{a\in\mathcal{A}}$=$\{3,3,4,4\} \rightarrow \ddot{y}_i$=$[1,1,1,0.5]$
- $\{y_{i,a}\}_{a\in\mathcal{A}}$=$\{0,1,2,2\} \rightarrow \ddot{y}_i$=$[0.75,0.5,0,0]$

If C=5, $|\mathcal{A}|$=3, and annotations,

- $\{y_{i,a}\}_{a\in\mathcal{A}}$=$\{0,1,1\} \rightarrow \ddot{y}_i$=$[0.66,0,0,0]$

B Datasets

Knee Osteoarthritis (KOA) dataset. The training set includes 2,286, 1,046, 1,516, 757, and 173 x-ray images for grades 0–4. The validation set contains 328, 153, 212, 106, and 27 respectively. We combine the above into a single training set for DisMix training. The testing split has 639, 296, 447, 223, and 51 respectively.

Indian Diabetic Retinopathy Image Dataset (IDRID). The training set contains 134, 20, 136, 74, and 49 fundus images for grades 0–4, respectively. The official test set includes 34, 5, 32, 19, and 13 images for grades 0–4.

Chaoyang dataset. The training set contains 1,111 normal, 842 serrated, 1,404 adenocarcinoma, and 664 adenoma images. The consensus-labeled test set includes 705, 321, 840, and 273 images, respectively.

Crowd Gleason dataset. The majority-vote training set contains 7,974 non-cancer, 2,038 grade 3, 2,840 grade 4, and 972 grade 5 images. The expert–resident consensus test set includes 2,157, 548, 164, and 57 images, respectively.

C Extended Results

C.1 Extended Analysis with Data Scarcity.

We extend the analysis under data scarcity to more datasets and methods in Tab. 9 and show that DisMix performs well under data scarce conditions.

C.2 Extended Analysis with Grading Variability.

We extend the comparison of grading variability levels and mix probabilities to more mixup baselines in 8. DisMix remains robust across grading variability levels of $\approx 30-70\%$ with mix probabilities of 0.3 and 0.5, outperforming the baselines in most settings, where each setting corresponds to a combination of dataset, ordinal classification method, grading variability level, and mix probability. A mix probability of 0.5 provides consistent performance as grading variability increases, including under severe variability ($\approx 70\%$). In contrast, increasing the mix probability to 0.7 generally reduces

Table 8: Extended performance comparison under different grading variability levels (30%, 50%, 70%) and mixing probabilities (0.3, 0.5, 0.7).

	IDRID				KOA				
	Accuracy ↑		MAE ↓		Accuracy ↑		MAE ↓		Mix
	ORD2SEQ	CLOC	ORD2SEQ	CLOC	ORD2SEQ	CLOC	ORD2SEQ	CLOC	Prob
≈30%									
GuidedMix	65.05 ± 0.97	66.65 ± 2.81	0.65 ± 0.04	0.57 ± 0.04	69.20 ± 0.50	67.87 ± 0.32	**0.41 ± 0.01**	0.42 ± 0.01	
DiffuseMix	64.73 ± 1.12	64.39 ± 1.12	0.66 ± 0.02	0.62 ± 0.06	69.50 ± 0.37	67.73 ± 0.31	0.42 ± 0.02	**0.41 ± 0.01**	0.3
DisMix	**66.02 ± 1.68**	**66.92 ± 0.13**	**0.61 ± 0.04**	**0.57 ± 0.01**	**69.57 ± 0.32**	**68.03 ± 0.82**	0.42 ± 0.02	0.42 ± 0.01	
GuidedMix	64.73 ± 1.48	64.07 ± 2.57	0.68 ± 0.03	0.61 ± 0.05	68.92 ± 0.31	67.57 ± 0.43	0.43 ± 0.03	0.42 ± 0.01	
DiffuseMix	65.70 ± 1.48	63.75 ± 1.12	0.62 ± 0.02	0.66 ± 0.02	68.68 ± 0.31	67.64 ± 0.98	0.43 ± 0.02	0.41 ± 0.01	0.5
DisMix	**66.34 ± 1.12**	**66.02 ± 0.98**	**0.58 ± 0.02**	**0.57 ± 0.02**	**69.41 ± 0.45**	**68.41 ± 0.36**	**0.40 ± 0.02**	**0.41 ± 0.01**	
GuidedMix	64.73 ± 2.02	**65.03 ± 3.37**	0.64 ± 0.03	**0.58 ± 0.07**	**69.26 ± 0.21**	67.77 ± 0.71	**0.41 ± 0.01**	0.42 ± 0.01	
DiffuseMix	**67.31 ± 0.56**	63.10 ± 2.57	**0.61 ± 0.01**	0.61 ± 0.05	69.02 ± 0.21	**68.21 ± 0.09**	0.43 ± 0.01	**0.41 ± 0.00**	0.7
DisMix	64.73 ± 1.48	64.73 ± 1.49	0.61 ± 0.04	0.64 ± 0.05	67.62 ± 0.77	68.09 ± 0.54	0.43 ± 0.01	0.41 ± 0.01	
≈50%									
GuidedMix	62.12 ± 1.70	64.39 ± 2.80	0.70 ± 0.02	0.59 ± 0.05	68.32 ± 0.19	67.93 ± 0.63	0.45 ± 0.02	0.42 ± 0.01	
DiffuseMix	**65.05 ± 1.94**	63.75 ± 2.24	**0.64 ± 0.01**	0.59 ± 0.05	68.38 ± 0.46	**67.99 ± 0.69**	0.45 ± 0.01	**0.41 ± 0.02**	0.3
DisMix	62.47 ± 1.13	**65.37 ± 1.49**	0.66 ± 0.06	**0.57 ± 0.03**	**68.46 ± 0.73**	67.93 ± 0.10	**0.45 ± 0.01**	0.42 ± 0.00	
GuidedMix	62.14 ± 0.97	**64.72 ± 2.02**	0.69 ± 0.01	**0.55 ± 0.07**	68.56 ± 0.27	68.07 ± 0.38	0.44 ± 0.03	0.41 ± 0.01	
DiffuseMix	63.76 ± 0.56	63.75 ± 1.12	0.71 ± 0.05	0.62 ± 0.04	**68.62 ± 0.33**	67.65 ± 0.70	0.46 ± 0.01	0.42 ± 0.01	0.5
DisMix	**66.44 ± 4.96**	62.78 ± 1.12	**0.63 ± 0.06**	0.61 ± 0.03	68.40 ± 0.35	**68.19 ± 0.50**	**0.44 ± 0.03**	**0.41 ± 0.01**	
GuidedMix	63.43 ± 0.56	62.78 ± 2.02	0.68 ± 0.04	0.61 ± 0.10	**68.74 ± 0.72**	67.08 ± 0.77	0.45 ± 0.01	0.43 ± 0.02	
DiffuseMix	63.11 ± 1.68	63.75 ± 2.24	0.72 ± 0.05	0.62 ± 0.02	68.52 ± 0.18	68.11 ± 0.32	0.46 ± 0.01	0.41 ± 0.01	0.7
DisMix	**64.09 ± 1.69**	**67.98 ± 3.34**	**0.64 ± 0.03**	**0.52 ± 0.02**	67.15 ± 0.65	**68.55 ± 0.46**	**0.44 ± 0.03**	**0.41 ± 0.01**	
≈70%									
GuidedMix	54.34 ± 2.58	56.92 ± 2.81	0.69 ± 0.00	0.70 ± 0.07	66.60 ± 1.37	**65.45 ± 2.21**	**0.48 ± 0.04**	**0.45 ± 0.04**	
DiffuseMix	55.01 ± 2.81	56.24 ± 3.35	0.70 ± 0.04	0.67 ± 0.15	66.20 ± 0.22	65.10 ± 1.86	0.49 ± 0.02	0.46 ± 0.03	0.3
DisMix	**55.35 ± 1.69**	**56.99 ± 0.59**	**0.68 ± 0.02**	**0.67 ± 0.01**	**66.81 ± 1.12**	64.46 ± 0.49	0.51 ± 0.05	0.47 ± 0.01	
GuidedMix	54.37 ± 0.97	56.63 ± 3.68	0.76 ± 0.04	0.72 ± 0.07	**67.03 ± 0.85**	63.91 ± 1.88	0.48 ± 0.03	0.46 ± 0.02	
DiffuseMix	54.66 ± 0.56	58.57 ± 5.35	0.73 ± 0.04	0.65 ± 0.05	65.88 ± 0.53	64.59 ± 1.30	0.50 ± 0.00	0.47 ± 0.01	0.5
DisMix	**55.02 ± 2.44**	**58.89 ± 6.46**	**0.71 ± 0.07**	**0.65 ± 0.03**	66.61 ± 0.58	**65.57 ± 0.31**	**0.48 ± 0.01**	**0.45 ± 0.01**	
GuidedMix	**57.00 ± 2.25**	**57.93 ± 2.80**	0.73 ± 0.02	**0.64 ± 0.06**	66.53 ± 1.48	**65.47 ± 1.36**	**0.48 ± 0.05**	**0.45 ± 0.02**	
DiffuseMix	56.94 ± 2.04	56.30 ± 1.69	**0.70 ± 0.05**	0.67 ± 0.07	**66.85 ± 0.06**	64.07 ± 0.79	0.51 ± 0.03	0.46 ± 0.01	0.7
DisMix	55.66 ± 3.12	55.33 ± 0.98	0.70 ± 0.07	0.72 ± 0.03	61.75 ± 1.18	64.85 ± 0.16	0.52 ± 0.03	0.46 ± 0.01	

Table 9: Extended results of performance under data scarcity, with the training set limited to 10% of samples per class ($\geq$10 images/class).

	Accuracy ↑		Mean Absolute Error ↓	
	ORD2SEQ	CLOC	ORD2SEQ	CLOC
IDRID				
Mixup	49.92 ± 1.06	43.36 ± 1.49	0.88 ± 0.05	1.06 ± 0.04
GuidedMix	49.82 ± 2.41	46.09 ± 1.77	0.88 ± 0.08	1.00 ± 0.09
DiffuseMix	48.22 ± 1.12	45.26 ± 2.84	0.93 ± 0.02	1.13 ± 0.08
DisMix	**51.44 ± 1.67**	**47.91 ± 2.82**	**0.86 ± 0.06**	**0.99 ± 0.09**
KOA				
Mixup	58.22 ± 1.19	58.77 ± 1.03	0.58 ± 0.02	0.57 ± 0.02
GuidedMix	58.01 ± 1.40	57.93 ± 0.07	0.60 ± 0.01	0.59 ± 0.03
DiffuseMix	58.13 ± 1.14	59.00 ± 0.98	0.59 ± 0.02	**0.56 ± 0.00**
DisMix	**58.58 ± 0.32**	**59.53 ± 1.40**	**0.58 ± 0.01**	0.58 ± 0.04
CHAOYANG				
Mixup	78.37 ± 0.20	79.52 ± 0.42	0.33 ± 0.01	0.31 ± 0.00
GuidedMix	78.99 ± 0.56	**80.02 ± 0.60**	0.33 ± 0.01	**0.30 ± 0.01**
DiffuseMix	78.20 ± 0.18	78.68 ± 0.74	0.33 ± 0.00	0.32 ± 0.02
DisMix	**79.38 ± 0.48**	78.77 ± 0.70	**0.32 ± 0.03**	0.32 ± 0.01
GLEASON				
Mixup	86.15 ± 1.19	84.82 ± 0.95	0.23 ± 0.01	0.21 ± 0.02
GuidedMix	86.35 ± 0.71	84.22 ± 0.22	0.21 ± 0.00	0.23 ± 0.01
DiffuseMix	87.07 ± 0.52	84.67 ± 1.46	0.23 ± 0.03	0.23 ± 0.01
DisMix	**87.54 ± 0.62**	**85.38 ± 0.76**	**0.20 ± 0.01**	**0.21 ± 0.01**

Table 10: Extended performance comparison of individual mixing policies.

	Accuracy ↑		Mean Absolute Error ↓	
	ORD2SEQ	CLOC	ORD2SEQ	CLOC
IDRID				
Ordinal Mix	66.67 ± 0.56	65.09 ± 1.73	0.61 ± 0.02	0.56 ± 0.02
Generate & Mix	64.73 ± 0.56	67.01 ± 2.56	0.62 ± 0.02	0.55 ± 0.04
Non-Ordinal Mix	64.73 ± 2.02	64.74 ± 1.17	0.59 ± 0.05	0.60 ± 0.01
Order Swap	65.37 ± 2.44	66.69 ± 2.96	0.65 ± 0.04	0.54 ± 0.06
KOA				
Ordinal Mix	68.90 ± 0.27	68.35 ± 0.79	0.43 ± 0.01	0.41 ± 0.01
Generate & Mix	68.72 ± 0.22	68.27 ± 0.40	0.42 ± 0.02	0.41 ± 0.01
Non-Ordinal Mix	68.36 ± 0.27	67.71 ± 0.33	0.42 ± 0.01	0.42 ± 0.01
Order Swap	68.88 ± 0.48	67.93 ± 0.36	0.43 ± 0.00	0.41 ± 0.00
CHAOYANG				
Ordinal Mix	84.29 ± 0.29	84.77 ± 0.39	0.27 ± 0.00	0.22 ± 0.01
Generate & Mix	84.14 ± 0.59	84.07 ± 0.71	0.26 ± 0.02	0.23 ± 0.02
Non-Ordinal Mix	84.10 ± 0.13	84.27 ± 0.88	0.25 ± 0.01	0.23 ± 0.01
Order Swap	83.54 ± 0.00	84.44 ± 1.11	0.26 ± 0.00	0.24 ± 0.03
GLEASON				
Ordinal Mix	92.12 ± 0.67	88.96 ± 1.31	0.23 ± 0.04	0.17 ± 0.01
Generate & Mix	91.89 ± 0.24	90.34 ± 1.43	0.27 ± 0.07	0.13 ± 0.02
Non-Ordinal Mix	91.81 ± 0.44	89.62 ± 1.19	0.26 ± 0.04	0.16 ± 0.02
Order Swap	92.32 ± 0.55	90.39 ± 0.97	0.29 ± 0.02	0.14 ± 0.02

the advantage of DisMix and can result in performance below the baselines, suggesting that excessive mixing amplifies the uncertainty introduced by grading variability.

C.3 Extended Analysis of DisMix Components and Policies.

We extend the analysis of DisMix components and policies to more datasets and ordinal methods in Tab. 10 and show that policies are competitive against each other but remains suboptimal to their combined use in Tab. 2.

C.4 Extended Analysis of Policy Behavior Under Extreme Conditions.

We examine policy behavior under severe data scarcity (10% per class) and extreme grading variability (70% of boundary grade variability) using CLOC on KOA dataset. As shown in Tab.12, Generate & Mix achieves $\approx 1-2\%$ higher accuracy than the respective combined-policies setting for CLOC

Table 11: DisMix performance with label-preserving vs. interpolating Mixups.

	Accuracy ↑		Mean Absolute Error ↓	
	ORD2SEQ	CLOC	ORD2SEQ	CLOC
IDRID				
Interpolating	**65.70 ± 2.02**	**65.36 ± 0.56**	0.65 ± 0.03	0.59 ± 0.02
Preserving	64.08 ± 2.57	65.04 ± 0.00	**0.62 ± 0.05**	**0.58 ± 0.02**
KOA				
Interpolating	68.56 ± 0.18	67.89 ± 0.25	0.41 ± 0.01	0.42 ± 0.00
Preserving	**68.62 ± 0.17**	**68.13 ± 0.41**	**0.41 ± 0.00**	**0.41 ± 0.00**
CHAOYANG				
Interpolating	83.79 ± 0.58	83.99 ± 0.59	0.28 ± 0.03	0.23 ± 0.01
Preserving	**83.82 ± 0.05**	**84.63 ± 0.41**	**0.27 ± 0.01**	**0.23 ± 0.00**
GLEASON				
Interpolating	91.37 ± 0.59	88.85 ± 0.89	**0.24 ± 0.09**	0.17 ± 0.02
Preserving	**91.65 ± 0.54**	**89.86 ± 0.34**	0.27 ± 0.05	**0.16 ± 0.00**

Table 12: Performance of mixup policies in severe data scarcity (10% data per class) and extreme grading variability (≈70%) with CLOC on KOA dataset.

	Severe Scarcity		Extreme Grading Variability	
	Accuracy ↑	MAE ↓	Accuracy ↑	MAE ↓
Ordinal Mix	59.75 ± 0.47	0.57 ± 0.00	64.45 ± 2.11	0.45 ± 0.02
Generate & Mix	**60.28 ± 1.13**	**0.56 ± 0.02**	**66.57 ± 2.11**	**0.44 ± 0.02**
Non-Ordinal Mix	59.60 ± 1.05	0.58 ± 0.02	65.27 ± 1.47	0.46 ± 0.02
Order Swap	*59.90 ± 0.27*	*0.57 ± 0.02*	*65.88 ± 0.07*	*0.45 ± 0.00*

Table 13: DisMix vs. base and fine-tuned (FT) DiffuseMix with CLOC on IDRID.

	Accuracy ↑	MAE ↓
Base	63.75	0.66
FT	64.30	0.61
DisMix	66.02	0.57

in Tab. 9 (with data scarcity) and Tab. 4 (70% grading variability with 0.5 mix probability), likely due to its two-stage process of generating a clean adjacent rank before mixing. This highlights the advantage of ordinal-conditioned synthesis over interpolation between available samples under extreme conditions. Order Swap, the second-best, indicates that label-preserving mixup remains effective even under extreme settings.

C.5 Extended Analysis of Label-Preserving vs. Interpolating Mixup.

Table 11 groups policies based on label-preserving (Non-ordinal Mix and Order Swap) and label-interpolating (Ordinal Mix and Generate & Mix) and show that label-preserving policies slightly outperform interpolating policies when used individually on KOA, Chaoyang and Gleason.

C.6 DisMix Training Overhead.

We compare the training overhead of DisMix and DiffuseMix on IDRID using a single RTX 4090 GPU. DisMix trains in only ≈1.3 hours, over 6× faster than DiffuseMix fine-tuning (≈8.5 hours), while achieving higher accuracy and lower MAE as shown in Tab. 13.